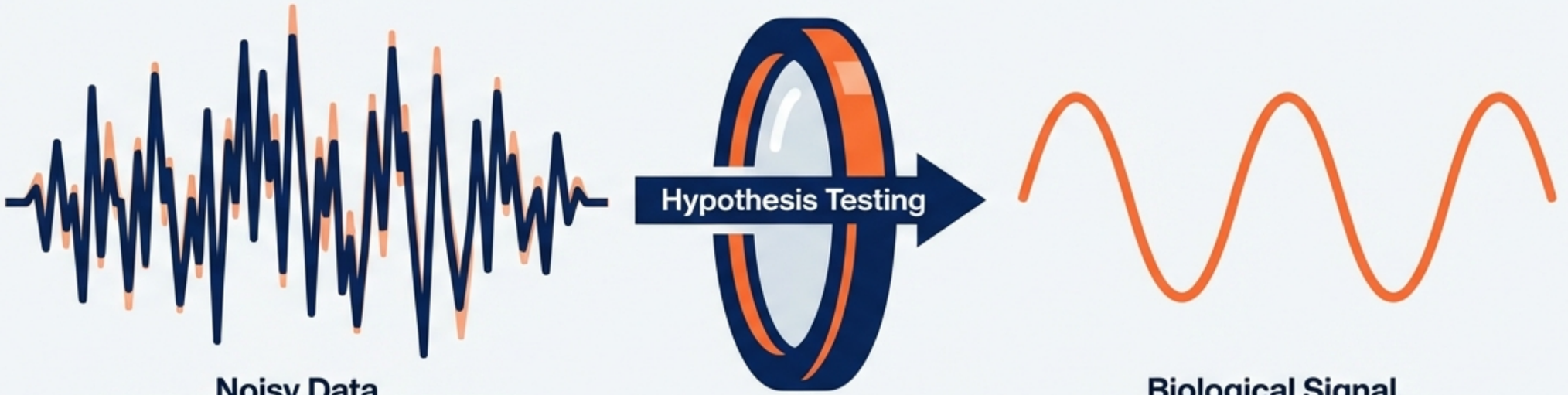
Hypothesis Testing and Statistical Inference

Makiing Principled Decisions Under Uncertainty in Movement Science

Hypothesis testing and statistical inference form the backbone of scientific discovery. While confidence intervals focus on estimation and uncertainty, this framework provides the necessary tools for making binary “yes/no” decisions about populations based on sample data. It is the method by which we determine if a training intervention truly improved performance, or if the results were merely a product of chance.

NotebookLM

The Core Challenge: Signal vs. Noise



In Movement Science, we constantly ask if an intervention—like a plyometric program—actually worked or if the observed changes were just random luck. Hypothesis testing provides a formal framework to answer this question. Unlike confidence intervals, which emphasize estimation, hypothesis testing focuses on a decision: we must choose between the “Null” status quo and the “Alternative” possibility of a real effect. This process, known as Null Hypothesis Significance Testing (NHST), has dominated scientific inference for nearly a century.

NotebookLM

Scientific proof relies on the logic of contradiction

The Null Hypothesis (H_0)

The “Presumption of Innocence.” It assumes the status quo: no effect, no difference, no relationship (e.g., $\mu = 50$ cm). We assume this is true until proven otherwise.

The Alternative Hypothesis (H_1)

The “Contradiction.” This is the claim of a real effect (e.g., $\mu \neq 50$ cm) that we hope to support. We only embrace H_1 if the evidence against H_0 is overwhelming.

We never “prove” the Null. We simply fail to reject it if the evidence is insufficient, much like a “not guilty” verdict does not prove innocence, but merely a lack of evidence for guilt.

NotebookLM

The p-value gauges the compatibility of of data with the Null

The p-value represents the probability of observing data as extreme as what we collected, assuming the Null Hypothesis is actually true.

A low p-value (typically < 0.05) suggests that the observed data would be incredibly rare if the Null were true, leading us to reject the Null.

Warning

Crucially, the p-value is NOT the probability that the Null Hypothesis is true. It is strictly a probability regarding the data.

NotebookLM

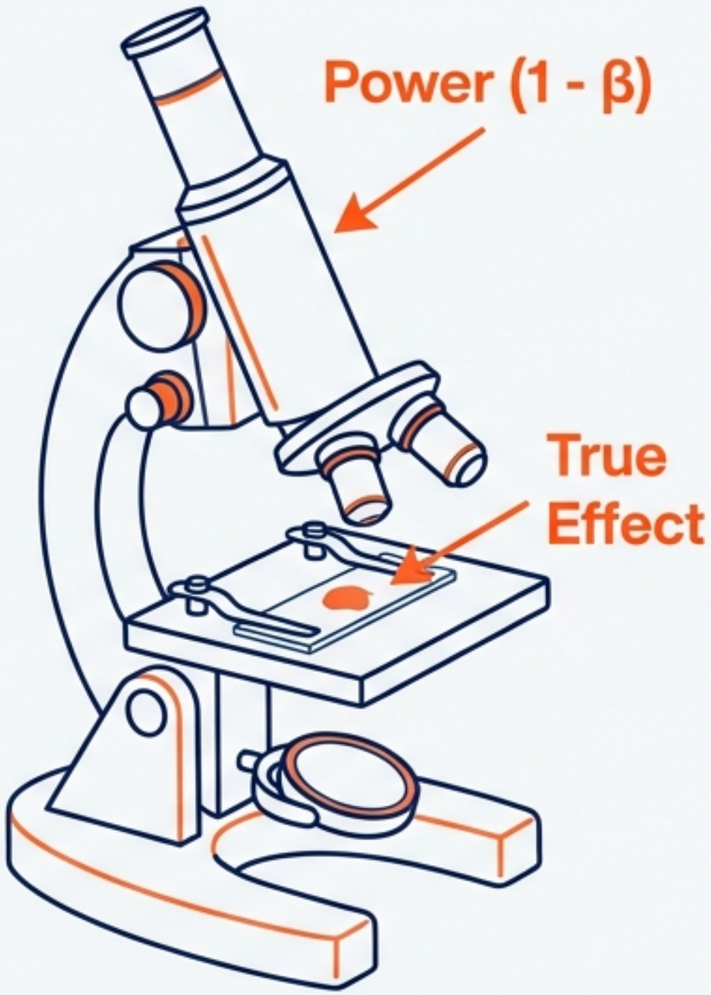
Relying on probability means accepting the risk of error

	Decision: Fail to Reject H_0	Decision: Reject H_0
Reality: H_0 is True (No Effect)	Correct Decision	Type I Error (α). False Positive. Concluding an effect exists when it doesn't.
Reality: H_0 is False (Real Effect)	Type II Error (β). False Negative. Missing a real effect that exists.	Correct Decision

Every statistical decision carries risk. A Type I error is a "false alarm," while a Type II error is a "missed discovery." Researchers typically accept a 5% risk of a False Positive ($\alpha = 0.05$). Reducing this risk inevitably increases the chance of missing a real effect.

NotebookLM

Statistical power is the ability to detect real effects



$$\text{Power} = f(\text{Sample Size } (n), \text{Effect Size, Alpha, Variability}).$$

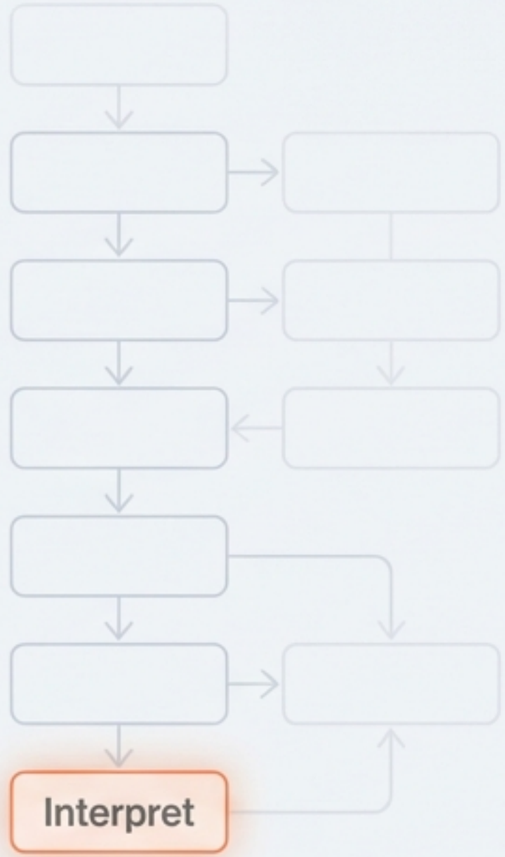
Statistical power is your study's "muscle"—the probability of correctly rejecting the Null when an effect truly exists. In Movement Science, underpowered studies are a major issue, often due to small sample sizes failing to detect meaningful training adaptations. Researchers should aim for a power of ≥ 0.80, meaning there is an 80% chance of finding the effect if it is there.

NotebookLM

A rigorous investigation follows a seven-step workflow

1. **State Question**
   Does plyometrics improve jump height?
2. **Formulate Hypotheses**
   Define H_0 and H_1.
3. **3. Set Significance**
   Choose α (typically 0.05).
4. **4. Select Test**
   Choose T-test based on design.

4. **4. Select Test**
   Choose T-test based on design.
5. **5. Compute**
   Calculate Statistic and p-value.
6. **6. Decide**
   Reject or Fail to Reject H_0.
7. **7. Interpret**
   Contextualize with Effect Size & CI.



Conducting a valid hypothesis test requires a disciplined sequence, not just a button press in software.

NotebookLM

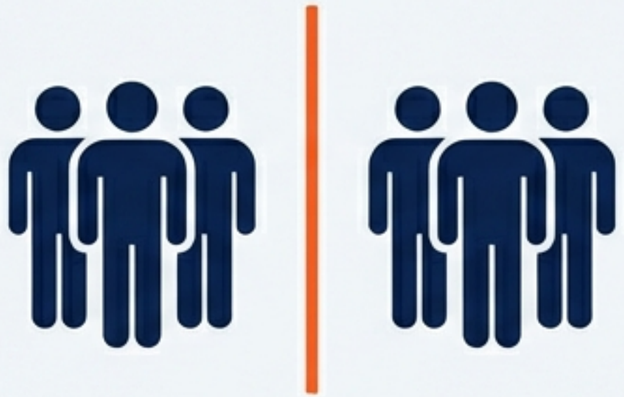
The t-test is the standard tool for comparing means

The specific tool depends entirely on the study design.

One-Sample

Comparing one group to a standard.

Independent Two-Sample



Comparing two distinct groups (e.g., Treatment vs Control).

Paired (Dependent)

Comparing the same subjects over time.

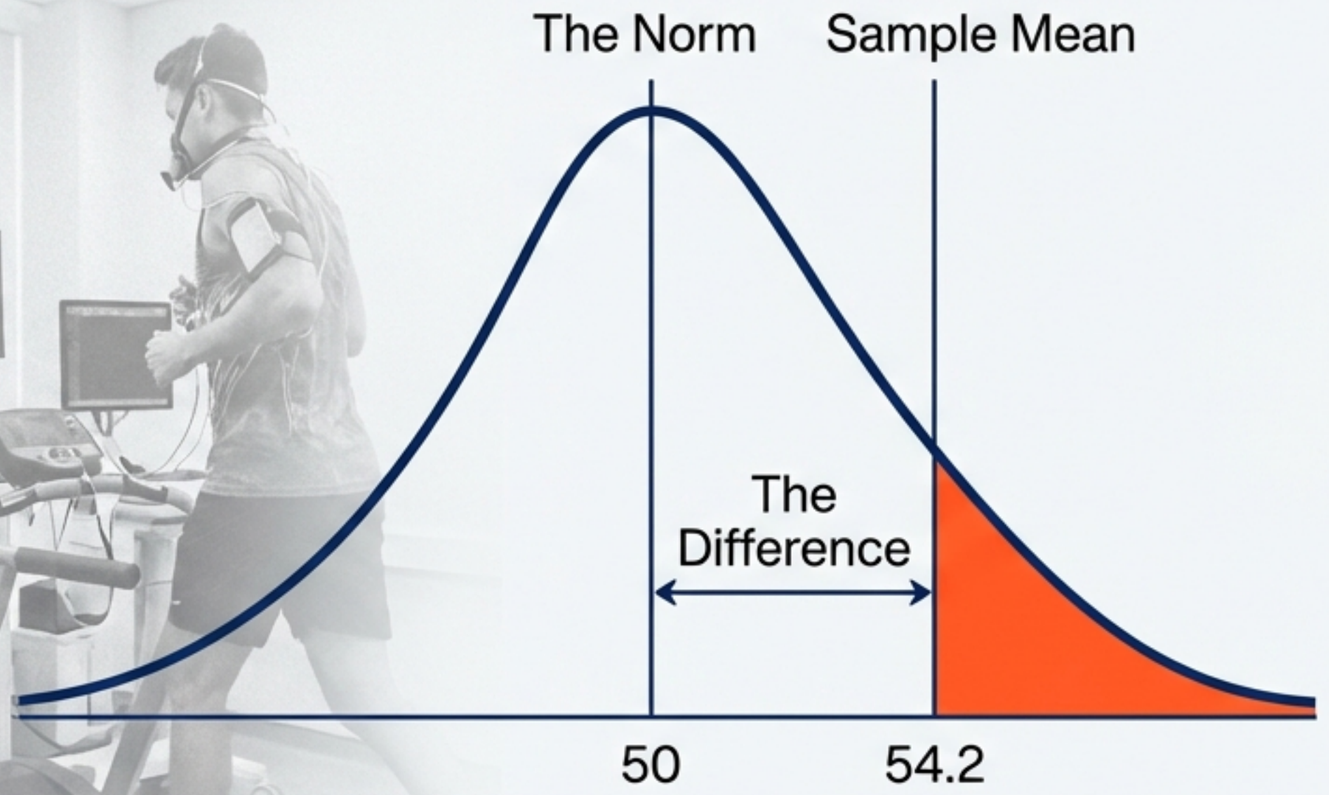
When the research question involves comparing means, the t-test is the workhorse. All t-tests rely on "degrees of freedom" (df) to account for independent information.

NotebookLM

Design 1: One-sample t-tests compare a group to a norm

Scenario: Comparing soccer team VO2max to a national norm of 50.

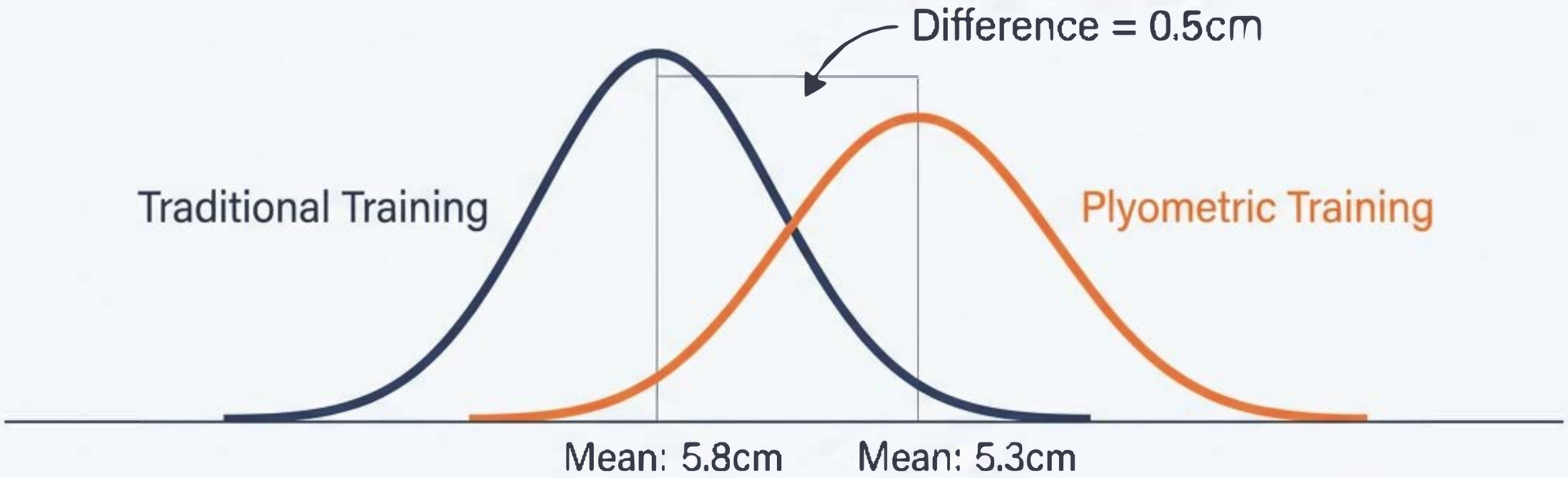
- **Sample Mean:** 54.2
- **Hypothesized Mean:** 50
- **Result:** t = 2.21, **p = 0.039**.



This test determines if a sample mean differs significantly from a known population value. If the calculated t-value exceeds the critical value defined by the degrees of freedom (n-1), we reject the Null.

NotebookLM

Design 2: Independent t-tests compare two distinct groups



Plyometric Group: 5.3cm | Traditional Group: 5.8cm

t(27.3) = 2.50, p = 0.019

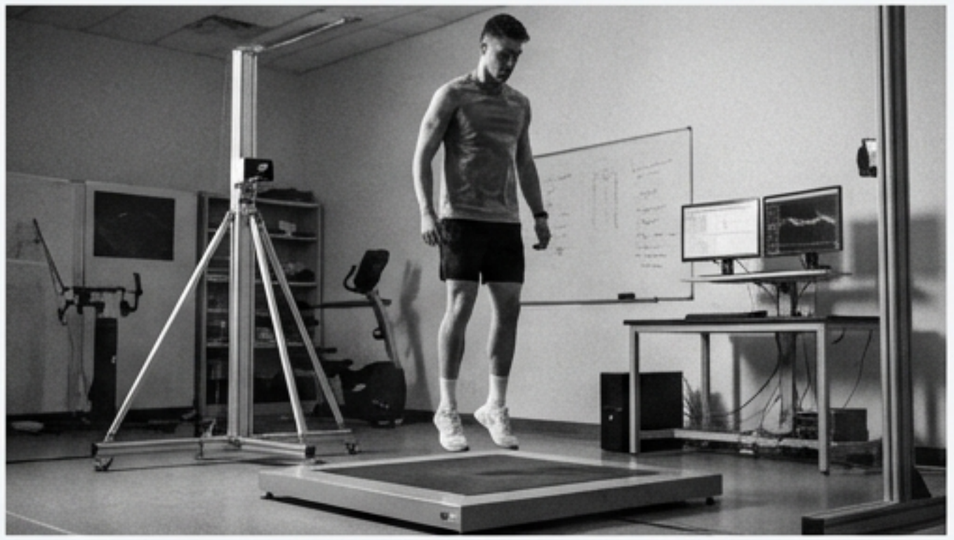
Conclusion: Significant Difference.

Used when comparing separate groups, such as a treatment vs. control. We check if the difference between their means is larger than what we'd expect by chance. We use the Welch approximation for degrees of freedom to account for unequal variances.

NotebookLM

Design 3: Paired t-tests monitor changes within subjects

When monitoring the same athletes over time (e.g., jump height before and after 8 weeks of training), the paired t-test analyzes the difference scores for each individual. This design is powerful because each participant acts as their own control, reducing noise from biological variability.



NotebookLM

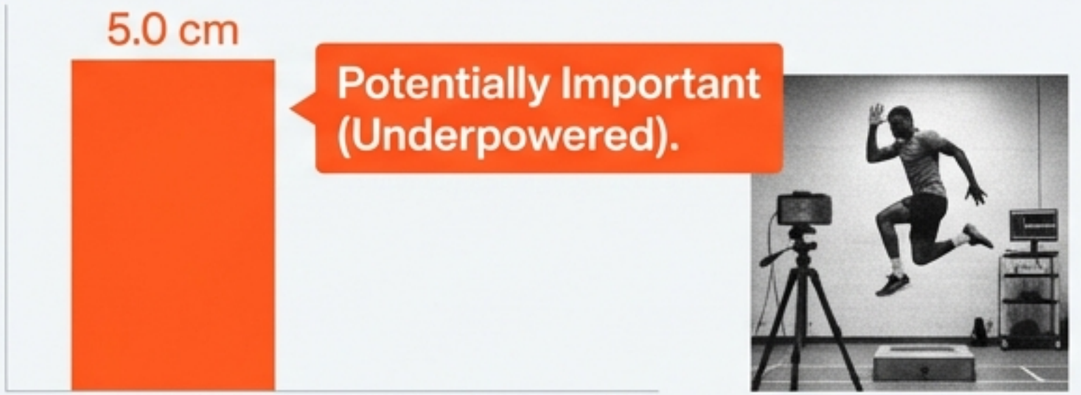
Statistical significance does not guarantee practical importance

Case A: The Massive Sample

n=500. Effect: 0.2 cm jump improvement.
p = 0.03 (Significant).

Case B: The Small Pilot

n=8. Effect: 5.0 cm jump improvement.
p = 0.08 (Not Significant).



A result can be **statistically significant** (unlikely to be chance) but practically useless. Conversely, a **meaningful effect might be missed due to low power.** Researchers must report Effect Sizes (like Cohen's *d*) to quantify the real-world magnitude of the result.

NotebookLM

Avoid common pitfalls when interpreting the verdict

DON'T

- **1. Don't be Binary.**

p=0.049 ≠ p=0.051

These are not opposite truths. Treat p-values as a gradation of evidence.

- **2. Don't Mistake Silence for Innocence.**

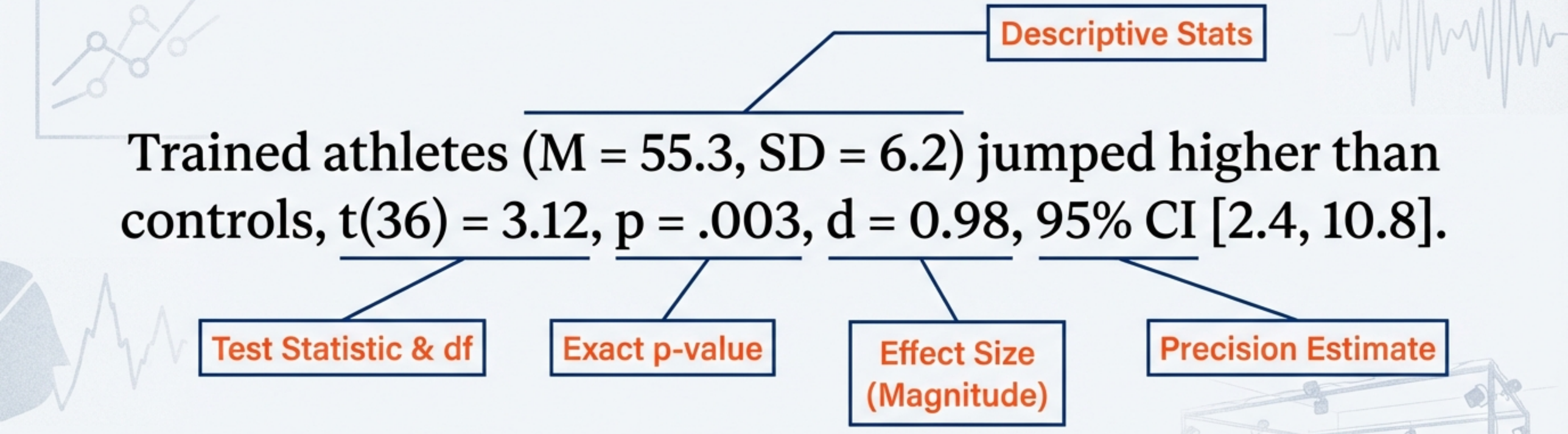
Failing to reject H_0 does not prove H_0 is true. It often just means we lack evidence (low power) to convict it.

- **3. Don't Ignore Context.**

Never interpret a non-significant result as "no effect" without checking the confidence confidence intervals.

NotebookLM

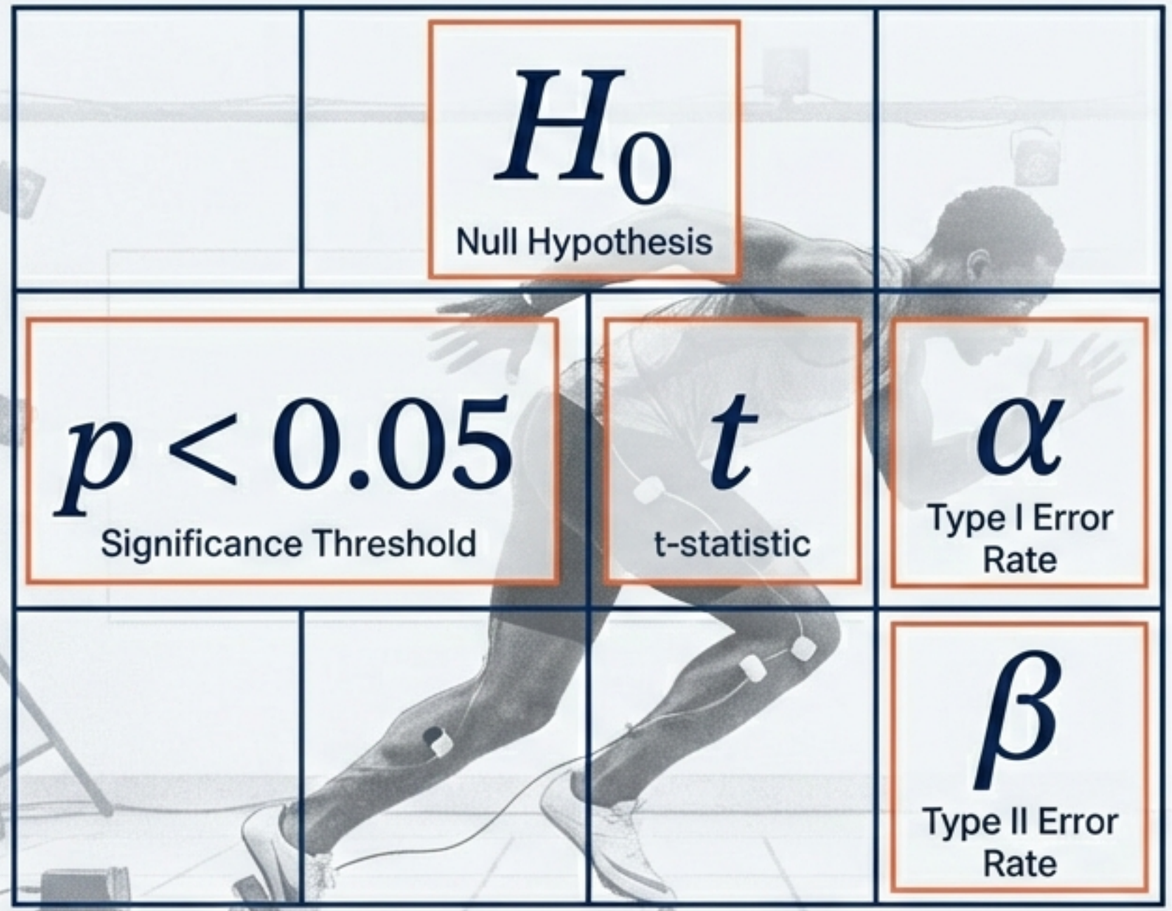
Transparent reporting requires more than just a p-value


Descriptive Stats
Trained athletes (M = 55.3, SD = 6.2) jumped higher than controls, t(36) = 3.12, p = .003, d = 0.98, 95% CI [2.4, 10.8].
Test Statistic & df
Exact p-value
Effect Size (Magnitude)
Precision Estimate

Transparency is key. Always report the descriptive stats, the specific test statistic, the exact p-value, and the effect size with confidence intervals.

NotebookLM

Hypothesis testing manages uncertainty in scientific discovery



Hypothesis testing is a rigorous framework for evaluating claims, not a machine for generating absolute truth. By formulating clear hypotheses, calculating test statistics, and carefully interpreting p-values alongside effect sizes, researchers can distinguish reproducible biological adaptations from random noise. We must balance these binary decisions with estimation to make the most informed inferences about human performance.

NotebookLM